# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. AGENCY USE ONLY (*Leave Blank*) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | 8 Jan 98 | Final, 1 Nov 93 – 31 Oct 97 |

4. TITLE AND SUBTITLE
General Coverage Problems with Applications, and Bootstrap Method in Survival Analysis and Reliability Theory

5. FUNDING NUMBERS
G
F49620-94-1-0035

6. AUTHORS
Shaw-Hwa Lo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Columbia University
Office of Projects and Grants
351 Engineering Terrace, MC 2205
NY, NY 10027

AFRL-SR-BL-TR-98-
0177

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

19980218 044

13. ABSTRACT *(Maximum 200 words)*

By approximating the classical Product-Limit estimator of a distribution function with an average of iid random variables, we derive, for the first time in the literature, sufficient and nessary conditions for the rates of (both strong as well as weak laws) uniform convergence of the Product-Limit estimator over the whole line. These findings somehow fill a longstanding gap in the theory of Survival Analysis and provide a systematic tool handling other challenging problems when data are incomplete. In deriving our main results we also suggested a heurestic way of estimating the rates of convergence. To demonstrate its applications , we prove a related conjecture of Gill and explain how a reliable confidence interval and band near the endpoint can be constructed.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| Product-Limit Estimator, Censored Data, Representations, Strong and Weak Laws, Stable Law, Martingale, Uniform convergence. | 16. PRICE CODE |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| unclassified | unclassified | unclassified | ul |

# Final Technical Report
# Under the AFOSR Grant F49620-94-1-0035
(November 1, 1993 - October 31, 1997)

**P.I.:** Shaw-Hwa Lo

**Institution:** Columbia University, Box 20, Low Memorial Library, NY, NY 10027

**Grant no:** F49620-94-1-0035

**Status of Effort:** During the past year I have been continuously working on the areas of Survival analysis, reliability theory, coverage problems, and the bootstrap method with applications. Some of these efforts were published (96-97) or to be published in the near future. See the publications list below. Currently I am working with my students on the problem of "adapt Cox models to cased-control study in Survival analysis". The method if developped will have numerous applications in seeking risk factors in various fields.

**Accomplishments / New findings:**

Some of my recent findings, under the AFOSR grant F49620-94-1-0035, on the areas of reliability and survival analysis are particularly encouraging. These results together with their implications are briefly described as follows:

It is well-known that the Kaplan-Meier (K-M) estimator often behaves very unstable and unreliable on the tail part of the survival function due to heavy censoring. The distributional properties responsibile for these behaviors are largely unclear. Although there have been some discussion along the line in recent literature (Gill (1983), Ying (1989), and Stute and Wang (1993)), the general problems of to what extent the valid inference can be drawn and what is the nature which dominates the convergence (or divergence) of the tail estimator, are virtually unknown and unanswered. The answers to these questions will not only add novel knowledge to the statistical literature, more importantly, they provide guidelines and suggest applicable procedures to the daily users routinely, as we shall see below.

Roughly speaking, our recent findings (some appeared in June, 1997 *Ann. Stat.*) indicate that the tail behaviors of K-M curve are determined by a set of simple necessary and sufficient conditions. There are two basic types of convergence involved here, the STRONG law and WEAK law (in probability). Therefore there are two conditions which determine K-M curve's behavior, depending on strong or weak laws respectively. If we denote by $\tau_H$ the right limit of support of $\bar{H} = \bar{F}\bar{G}$, where $\bar{F}$ and $\bar{G}$ are usual survival function of interest and censoring survival function, respectively, the K-M estimator can be defined as

$$\hat{F}_n(t) = 1 - \prod_{0 \le s \le t} (1 - \frac{dN_n(s)}{Y_n(s)}),$$

where $N_n(s)$, and $Y_n(s)$ are the usual counting processes based on complete data and risk

size at remaining time $s$, respectively.

We outline major findings in the following RESULTS: The first result gives the necessary and sufficient condition of rates convergence in strong law.

**RESULT 1.** (STRONG LAW)

*For any* $0 \leq p \leq \frac{1}{2}$, *we conclude that*

$$\sup_{t \leq \tau_H} n^p |\hat{F}_n(t) - F(t)| = o(1) \quad a.s.$$

*if and only if*

$$\int_0^{\tau_H} (1-G)^{\frac{-p}{1-p}} \, dF < \infty. \tag{1}$$

The second result extents Lo and Singh's (1986) iid representations from finite interval to whole real line. Note that the iid variables $\{\xi\}$ appeared in the following expression are different from that in Lo and Singh (1986).

**RESULT 2.** (REPRESENTATIONS OF $\hat{F}_n - F$)

*If (1) holds for some* $p, 0 \leq p \leq \frac{1}{2}$, *then*

$$\hat{F}_n(t) - F(t) = \frac{\bar{F}(t)}{n} \sum_{i=1}^{n} \xi(Z_i, \delta_i, t) + \gamma_n(t)$$

*where* $\xi(z_i, \delta_i, t) = \frac{\delta_i I(z_i \leq t)}{1-H(z_i-)} - \int_0^t \frac{dH_1}{1-H}$, *and*

$$\sup_{t \leq \tau_H} |\gamma_n(t)| = 0(n^{-\frac{1}{2}} \vee n^{-\frac{3p}{2}})(\log n)^{\frac{1}{2}}) \quad a.s.$$

The following result tells us the rates of week convergence, which is of fundamental importance in statistics.

**RESULT 3** ( WEAK LAW)

*Assuming that* $0 \leq p \leq \frac{1}{2}$. *Then*

$$\sup_{t \leq \tau_H} |n^p(\hat{F}_n(t) - F(t)| = o_p(1) \qquad \textit{if and only if}$$

$$\overline{\lim}_{t \to \tau_H} \frac{\{\int_t^{\tau_H} (1-G) dF\}^{1-p}}{1-G(t)} < \infty.$$

*Furthermore, we conclude that*

$$\sup_{t\le\tau_H} |n^p(\hat{F}_n(t) - F(t)| = o_p(1) \qquad \textit{if and only if}$$

$$\lim_{t\to\tau_H} \frac{\{\int_t^{\tau_H}(1-G)dF\}^{1-p}}{1-G(t)} = 0.$$

A nature question arises: Does exact order of weak convergence exit? If it does, how to find it?

The following result 4 and 5 answer the question affirmatively.

**RESULT 4.**

*There exists a unique* $p$, $0 \le p \le \frac{1}{2}$ *such that*

$$\lim_{m\to\infty} \lim_{n} P\{0 < \sup_{t\le\tau_H} |n^p(\hat{F}_n(t) - F(t))| < m\} = 1.$$

This result also tells us that for that unique $p$, the following inequalities must hold:

$$\begin{aligned} -\infty &< \underline{\lim}_{t\to\tau_H}(\log(1-H(t)) - (1-p)\log(\int_t^{\tau_H}(1-G)dF)) \\ &\le \overline{\lim}_{t\to\tau_H}(\log(1-H(t)) - (1-p)\log(\int_t^{\tau_H}(1-G)dF)) \\ &< \infty. \end{aligned}$$

Therefore, we can consider a nature estimator of $p$ as follows:

**RESULTS 5.** (ESTIMATION OF $p$)

*Consider a simple linear regression problem with observations* $\{(\log j, \log(N_n(Z_{(n-j)}))\}$, *where* $\{Z_{(j)}\}$ *are the ordered values of* $\{Z_i\}$. *If we treat* $\log j$ *as the covariate and* $\log(N_n(Z_{(n-j)}))$ *as the response variable in the linear model, then* $p$ *can be estimated consistently by* $\hat{p}$, *where* $1-\hat{p}$ *is the slop estimator described by the above regression problem.*

Although the proposed estiamtor $\hat{p}$ is consistent, the distributional properties of $\hat{p}$ is not clear at the moment. Further study is needed toward this important direction.

Another question needs to be explored is that from Result 4 above, there exists a unique $p$ such that $n^p(\hat{F}_n(t) - F(t))$ forms a tight sequences of processes on $D[0, \tau_H]$. The distribution of this limiting process is not identified, however. One may use bootstrap to approximate the limiting distribution in practice without knowing the theoretical distribution, the trouble is that the current case is not regular and the well known bootstrap theory cannot be applied.

The phenomenon discovered here, we believe, is both practically and theoretically important. To our best knowledge, nothing of this kind has been reported in the statistical literature. We also believe that the above phenomenon is not isolated. As a matter of fact, we expect that similar phenomenon will also occur in other incomplete data problems. Therefore we strongly believe that the techniques developed here can be employed to explore further important unknown.

**Progress on the area of**

## THE CASE-CONTROL STUDY WITH FAILURE TIME DATA

In most case-control studies, it is necessary to define a specific population where the cases and controls are randomly selected from the individuals who developed disease in a specificed accession period or from the individuals who are disease-free by the end of the case accession period, respectively. *Our situation is slightly different, however. The disease cases are randomly selected from a time-dependent population* $P_1(\tau_0)$ *consisting of individuals who meet certain inclusion criteria set before the study and are known to have developed disease before current calendar time* $\tau = \tau_0$. For example, suppose we are interested in a lung cancer study with risk variables $\underset{\sim}{X}$. The current calendar time is 1996 which is $\tau_0$. The inclusion criteria may include: (1) individuals who reside at certain areas, (2) he (or she) must be at least 20 years old but not exceed 80 by the present time $\tau_0$. An individual who lived in the specified area, now 48 and developed lung cancer 6 years ago (regardless dead or still alive) would be considered as a sample point in $P_1(\tau_0)$. The *survival time* $T$ of interest for this individual is clearly 22 years (since 42-20=22). An individual who was diagnosed lung cancer in 1962 at age of 53 is not included in $P_1(\tau_0)$, however.

Similarly, the controls sample are randomly selected from a time-dependent population $P_0(\tau_0)$ consisting of those individuals who meet certain similar inclusion criteria and known to be disease-free at $\tau_0$. In the above example, an individual who is 35 years and is disease-free contributes at least 15 years to the survival time $T$ of interest. In other words, $T$ is right censored. In fact, every individual's survival times $T$ in $P_0(\tau_0)$ is censored. This is why the individual is qualified to serve as a control.

Note that some of the cases in $P_1(\tau_0)$ may be censored (either left or right) for various reasons. For a clear and simple illustration of our proposed method, we have chosen to exclude this possibility that $P_1(\tau_0)$ may have incomplete observations. But the method proposed here does extend to cover these more complicated situations.

We shall assume that $P_1(\tau_0)$ consists of all units with complete survival time $\{T_\alpha, \alpha \epsilon P_1(\tau_0)\}$ while $P_0(\tau_0)$ consists of all individuals with right censored survival times $\{C_\gamma; \gamma \epsilon P_0(\tau_0)\}$ with $C_\gamma < T_\gamma$ for each $\gamma \epsilon P_0(\tau_0)$. The method considered here assumes that a suitable time-dependent population $\{P_1(\tau), P_0(\tau); \tau = \text{calendar time}\}$ can be defined. Some individuals may fall in $P_0(\tau)$ initially and subsequently develop the disease and become part of $P_1(\tau')$ for $\tau' > \tau$. For each individual with risk variables $\underset{\sim}{X}(t)$ (this $\tau$ corresponds to

survival time which is different from $t$), the disease incidence rate is defined as, according to Cox proportional hazards models,

$$\lambda(T = t|\underset{\sim}{X}(t)) = \lambda_0(t) \exp\{\underset{\sim}{X}(t)\underset{\sim}{\beta}^{\mathrm{T}}\}.$$

From now on we shall restrict our discussion on the case that risk variables are independent of time $t$. We shall return and comment on this general case later. It is plausible to estimate $\underset{\sim}{\beta}$ but not $\lambda_0(t)$ based on the case-control data. Suppose that $n$ cases and $m$ controls are *randomly* selected from $P_1(\tau_0)$ and $P_0(\tau_0)$ respectively. Let $Z = 1$ if someone is included in the sample, $Z = 0$, otherwise. (Note that the random indicator variable $Z$ defined here may depends on $\tau_0$).

Suppose that $n$ cases and $m$ controls are randomly selected from $P_1(\tau_0)$ and $P_0(\tau_0)$ seperately. Let $Z = 1$ if someone is included in the sample, $Z = 0$ otherwise. (Note that $Z$ may depend on $\tau_0$ and $\underset{\sim}{X}$, but given the disease status (i.e.; either in $P_1(\tau_0)$ or $P_0(\tau_0)$, $Z$ is independent of $\underset{\sim}{X}$ at $\tau = \tau_0$, however).

The full likelihood based on the observed data can be written as

$$\begin{aligned} Lik &= \prod_{i=1}^{n} f(T_i = t_i, \underset{\sim}{X}_i | T_i < C_i) \prod_{j=1}^{m} P(T_j > c_j, C_j = C_j, \underset{\sim}{X}_j | T_j > C_j) \\ &= L_1 \times L_2, \qquad \text{say} \end{aligned}$$

$L_1$ can be further written as

$$L_1 = \prod_{i=1}^{n} f(T_i = t_i, \underset{\sim}{X}_i | Z_i = 1, T_i < C_i),$$

since given $T < C, T$ and $\underset{\sim}{X}$ are independent of $Z$.

We then arrive at

$$\begin{aligned} L_1 &= \prod_{i=1}^{n} f(T_i = t_i | Z_i = 1, T_i < C_i) P(\underset{\sim}{X}_i | T_i = t_i, Z_i = 1, T_i < C_i) \\ &= \prod_{i=1}^{n} f(T_i = t_i, T_i < C_i | \underset{\sim}{X}_i, Z_i = 1) \frac{P(\underset{\sim}{X}_i | Z_i = 1)}{P(T_i < C_i | Z_i = 1)}. \end{aligned}$$

It is clear from the sampling plan, $P(T < C|Z = 1) = \frac{n}{n+m}$. Likewise, $L_2$ can be written as

$$L_2 = \prod_{j=1}^{m} P(T \geq c_j, T_j \geq C_j = C_j | \underset{\sim}{X}_j, Z_j = 1) \frac{P(Z_j = 1 | \underset{\sim}{X}_j)}{P(T_j < C_j | Z_j = 1)}.$$

and $P(T > C|Z = 1) = \frac{m}{n+m}$. We attempt to maximize $L_1 \times L_2$, subject to the following constraints

$$\frac{n}{n+m} = \sum_{\{\underset{\sim}{x}\}} P(T \leq C | \underset{\sim}{X} = \underset{\sim}{x}, Z = 1) P(\underset{\sim}{X} = \underset{\sim}{x} | Z = 1)$$

$$\frac{m}{n+m} = \sum_{\{\underset{\sim}{x}\}} P(T > C|\underset{\sim}{X} = \underset{\sim}{x}, Z = 1)P(\underset{\sim}{X} = \underset{\sim}{x}|Z = 1),$$

the summation runs over all possible exposure values, and will be replaced by integral if $\underset{\sim}{X}$ is continuous. An argument similar to Anderson (1972) and Prentice and Pyke (1979) shows that this constrained maximum likelihood estimate (MLE) is the same as the unconstrained MLE which maximizes

$$\prod_{i=1}^{n} f(T_i = t_i, T_i < C_i|\underset{\sim}{X}_i, Z_i = 1) \prod_{j=1}^{m} P(T_j \geq c_j, C_j = c_j|\underset{\sim}{X}_j, Z_j = 1) = L_1^* \times L_2^*, \text{ say.}$$

Now $L_1^*$ is proportional to $L_1^{**} = \prod_{i=1}^{n} f(T_i = t_i, T_i < C_i|\underset{\sim}{X}_i)$ and $L_2^*$ is proportional to $L_2^{**} = \prod_{j=1}^{m} P(T_j > c_j, C_j = c_j|\underset{\sim}{X}_j)$. This *says that if the prospective Cox model were applied to the case-control data, as if the sampling were prospective, the likelihood function would be proportional to the prospective likelihood. Therefore one can estimate the parameters* $\underset{\sim}{\beta}$ *exactly the same as the ordinary propspective partial likelihood method. The major difference is, the base line hazard function* $\lambda_0(t)$ *and corresponding cumulative hazard* $\Lambda(t) = \int_0^t \lambda_0(u)du$ *are no longer estimable, as evidenced by C.2 and the fact that "although* $L_1^{**} \times L_2^{**}$ *is proportional to* $L_1^* \times L_2^*$*, as demonstrated above, the constant factor (ratio) between the two products depends on the quantities such as* $P(Z = 1|T = t, T \leq C, \underset{\sim}{X})$ *and* $P(Z = 1|T > c, T > c, \underset{\sim}{X})$ *which are not estimable under the case-control design".* These two quantities here involve the knowledge of the size of $P_1(\tau_0)$ and $P_0(\tau_0)$, which is generally non-existent based on a case-control design, unless additional sources of information are available.

**FURTHER WORK.** The distributional properties of the estimator $\hat{\underset{\sim}{\beta}}$ derived above is important. We proposed to explore this fully in the near future. The issues of how to extend our method to accommodate more general case-control studies which involve continuous risk variables and time-dependent exposure are certainly interesting and important. With minor modifications of our method we feel we do can handle the time-dependent cases without much difficulties. To cover the general cases involving arbitrary continuous covariates will, however, require smoothing techniques and special treatment, and these deserved further study. We plan to include these problems in our future study.

So far we have discussed our method under the simplest design for case-control data. We believe we can do various extensions: Various degrees of matching or stratification can be built into the design and case-control sampling fractions can be allowed to vary among marked sets or stratus. The later issues is particular relevant to the problem raised earlier: what if we collect another case-control data 5 years from now? Can we combine these two data sets collected in distinct calendar times to conduct a coherent analysis? The answer to this question is "yes" under the simple design described above. Although the sampling probabilities may vary, depending on distinct calendar times and the corresponding time-dependent populations $\{P_1(\tau)\}$ and $\{P_0(\tau)\}$, the likelihood function can still be derived and shown to be proportional to the likelihood function derived from a prospective study. It

is not clear to us, however, whether this desirable property still holds in more complicated designs which involve matching and stratifications. This together with various practical variations of the problems constitute a further area of our future study. Furthermore, to test our method, we plan to apply our method to various existing data collected from earlier cancer studies.

## Publications (including to be published)

1. Maximum likelihood summary and the bootstrap mehtod in structured finite populations. (1994), *Statistica Sinica* Vol. 4, NO. 2, pp. 389-406. (with Min-Te Chao)

2. Some statistical mean value theorems related to the bootstrap. (1995), *Statistica Sinica* Vol. 5, No. 1, pp. 129-139. (with Min-Te Chao)

3. On strong uniform consistency of the Lynden-Bell estimator for the truncated data. (1995) *Annals Statistics* Vol. 23, No. 2, pp. 440-49.

4. On bootstrap accuracy with censored data. (1996), *Annals Statistics* vol. 24, No. 2, 569-595. (with K. Chen)

5. On a mapping approach to investigating the bootstrap. (1996) *Prob. Theory and Related Fields* , 107, 197-217. (with K. Chen)

6. On the rates of convergence of product-limit estimator over the whole real line: weak and strong laws. (1997) *Annals Statistics*, vol. 25, no. 3, 1050-1087. (with K. Chen)

7. Fitting stochastic abundance models via the Zipf's law: Another look at the Shakespeare data. (1997) *Biometrics.* Under review. (with M-T Chao and J-S Hwang)

## Personnel supports and Ph.D. thesis sponsorships

The following three Ph.D. students have been supported by this AFOSR grant during the preparation of their dissertations.

1. Kani Chen, Ph.D. June 1994.

   Dissertation: The bootstrap accuracy: a general mapping approach and the Edgeworth expansion with censored data.

2. Lihong Li, Ph.D. June, 1995.

   Dissertation: Almost sure representations and two sample problems with left truncated and right censored data.

3. Xiaofeng Gu, Ph.D. June, 1995.

   Dissertation: Estimating the treatment effect for the two-sample problem with truncated data.

**New discoveries:** none

**Honors / Awards:**

Shaw-Hwa Lo is a full professor of Statistics and Biostatistics. He is also now Co-chairman of Statistics Department at Columbia University, a Fellow of IMS and a Fellow of ASA.